УДК 004.075

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ПРОЦЕСУ ОБСЛУГОВУВАННЯ НАВАНТАЖЕННЯ В ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНІЙ МЕРЕЖІ

Н.А. Прокопець, Л.С. Глоба
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

Abstract *– The functioning of a modern information and communication network is impossible without software, which partially replaces specialized equipment when solving network tasks. These tasks form a workload that needs to be processed by a distributed server infrastructure within the network. There are requirements for each type of workload regarding the quality of its processing (maximum service time, permissible losses, etc.). Thus, there is a need to establish relationships between quality indicators and parameters of the processing system to ensure the fulfillment of these requirements using mathematical modeling. The paper proposes a mathematical model of a distributed workload processing system in the information and communication network as a queuing system. Unlike known models, the proposed model considers the possible variable nature of the input workload arrival rate and parallelization methods that may be applied to the software that implements it. Within the modeling process, a method of transition from a non-stationary non-ordinary incoming queries' flow to an ordinary stationary flow is proposed. Based on the constructed mathematical model, a complex method of energy-efficient workload processing has been developed. A laboratory experiment has proven the efficiency of the proposed complex method and the mathematical model's adequacy in its basis.*

Анотація *– Функціонування сучасної інформаційно-комунікаційної мережі неможливе без програмного забезпечення, яке частково замінює спеціалізоване обладнання при вирішенні мережних задач. Ці задачі формують навантаження, що потребує обслуговування за допомогою розподіленої серверної інфраструктури у складі мережі. Для кожного типу навантаження висуваються вимоги щодо якості його обслуговування (максимального часу обслуговування, допустимих втрат тощо). Для забезпечення виконання цих вимог виникає необхідність встановлення взаємозв'язків між показниками якості та параметрами системи обслуговування за допомогою математичного моделювання. У роботі запропоновано математичну модель системи розподіленого обслуговування навантаження в інформаційно-комунікаційній мережі у вигляді системи масового обслуговування. На відміну від відомих моделей, запропонована модель враховує можливий змінний характер інтенсивності вхідного навантаження та можливість паралелізації програмного забезпечення, що його реалізує. У процесі моделювання запропоновано метод переходу від нестаціонарного неординарного вхідного потоку заявок до стаціонарного ординарного потоку. На основі побудованої математичної моделі розроблено комплексний метод енергоефективного обслуговування навантаження. Ефективність запропонованого комплексного методу та адекватність математичної моделі в його основі доведено шляхом лабораторного експерименту.*

Вступ

У сучасних інформаційно-комунікаційних мережах (ІКМ) левова частка функцій технічних засобів реалізується за допомогою програмного забезпечення (ПЗ), що обумовлено розвитком ряду технологій і концепцій, зокрема SDN (Software-Defined Networking) [1], NFV (Network Functions Virtualization) [2], логічного поділу мережі (Network Slicing) [3], периферійних обчислень (Edge Computing) [4] та bDDN (Big data driven networking) [5]. У межах цих концепцій за допомогою ПЗ вирішуються, наприклад, такі завдання: маршрутизація, комутація, перетворення мережних адрес в NFV [2]; оркестрація та реконфігурація мережних ресурсів в SDN [1]; виокремлення логічних рівнів мережі та аналіз вимог сервісів на кожному з них у Network Slicing [6]; локальне кешування контенту для користувацьких додатків, обробка даних систем Інтернету речей чи відеоспостереження в Edge Computing[7]; аналіз даних про поточну конфігурацію мережі, стан мережного обладнання тощо в bDDN [5]. Ці

задачі складають обчислювальне навантаження, водночас особливості кожного типу навантаження формують специфічні вимоги щодо його обслуговування. Це зокрема вимоги щодо забезпечення якості обслуговування (Quality of Service, QoS) для кінцевого користувача, гарантованої доступності системи (ймовірності втрати запиту не вище визначеного рівня), максимально допустимого часу відповіді на запит, енергоефективності, безпеки тощо.

Обслуговування навантаження здійснюється за допомогою розподілених обчислювальних систем, що стали невід'ємною частиною архітектури сучасних ІКМ [8]. Такі системи фізично представлені центрами обробки даних, що складаються із серверних кластерів. Кожен серверний кластер зі свого боку складається із окремих серверів, що у загальному випадку можуть різнитись за своїми технічними параметрами (мати різну продуктивність, різну кількість обчислювальних ядер тощо). Задачі ІКМ у загальному випадку можуть виконуватись за допомогою ПЗ, реалізованого із використанням технік паралелізації, тобто використовувати більше одного обчислювального ядра для обробки [9]. З метою забезпечення виконання вимог щодо обслуговування різних типів навантаження ІКМ виникає необхідність створення математичної моделі процесу розподіленого обслуговування навантаження, яка дозволила б формально описати взаємозв'язки між параметрами системи обслуговування та показниками якості обслуговування, щодо яких висуваються вимоги.

І. Аналіз існуючих математичних моделей систем розподіленого обслуговування навантаження

На сьогодні існує низка підходів щодо вирішення проблеми математичного моделювання процесу обслуговування навантаження у розподіленій обчислювальній системі. Авторами роботи [10] була запропонована модель процесу обслуговування навантаження у гетерогенному центрі обробки даних (ЦОД) як системи масового обслуговування (СМО). Автори пропонують представити систему у вигляді двох СМО, що взаємодіють між собою. Основним недоліком такої моделі є те, що автори не розглядають можливість обробки одного запита користувача на багатьох процесорах, тобто не розглядають можливість паралелізації задач.

У роботі [11] запропоновано модель кластера зі скінченної кількості гетерогенних серверів, що мають різну інтенсивність обробки запитів як СМО. У роботі також розглядаються окремі черги до кожного з серверів, при чому довжина цих черг у запропонованій СМО є скінченою. Основним недоліком СМО у дослідженні [11] є те, що автори розглядають лише вхідний пуассонівський потік запитів зі сталою інтенсивністю. Для обчислювальних робіт у межах навантаження ІКМ інтенсивність надходження зазвичай є змінною та представлена випадковою величиною. Крім того, автори роботи також не розглядають можливість паралелізації задач, а отже напряму ставлять запит СМО у відповідність обчислювальній роботі, що у випадку застосування паралелізації може не відповідати дійсності.

Автори дослідження [12] запропонували модель серверного кластера як СМО, у якій навантаженням виступають запити користувачів. Вхідне навантаження розподіляється між окремими серверами за допомогою брокера планування навантаження. Автори ввели припущення, що якщо швидкість надходження запитів у кластер дорівнює λ, то швидкість надходження на робочий сервер кластера дорівнює λ_{p_C}, де p_C представляє пропорцію потужності сервера до загальної потужності системи. На кожному із серверів додатково формується окрема черга запитів. Недоліком даного підходу є те, що автори напряму зіставляють запити користувачів із запитами СМО, що значно спрощує процес. Адже у реальній системі один запит користувача може потребувати для своєї обробки більш ніж одне обчислювальне ядро сервера, а сервери зі свого боку можуть мати різну кількість обчислювальних ядер.

Автори роботи [13] замість представлення СМО у вигляді двох частин запропонували багатоканальну модель із пуассонівським вхідним потоком, що може бути розділена на *k* одноканальних систем шляхом декомпозиції. Особливо цікавим у цій роботі є те, що автори врахували можливість зміни інтенсивності обслуговування кожного окремого сервера внаслідок застосування апаратної техніки енергозбереження DVFS, яка застосовується практично в усіх сучасних процесорах [14]. Однак інтенсивність надходження запитів вважається сталою, що у загальному випадку не відповідає дійсності для навантаження ІКМ.

Таким чином, існуючі моделі не враховують того, що інтенсивність надходження запитів для навантаження ІКМ у загальному випадку не є сталою. Крім того, існуючі дослідження не враховують факту, що при розробці сучасного ПЗ часто застосовуються техніки паралелізації, що призводить до того, що одна обчислювальна робота може вимагати для свого обслуговування $k > 1$ обчислювальних ядер. Метою даного дослідження є створення математичної моделі системи розподіленого обслуговування навантаження ІКМ, що враховує змінний характер інтенсивності навантаження, а також можливість використання технік паралелізації при розробці ПЗ сучасних мереж.

II. Математична модель досліджуваної системи

Розглянемо розподілену обчислювальну систему, що складається з *M* серверних кластерів, кожен з яких зі свого боку складається з *N* обчислювальних вузлів. Кожен *j*-й вузол характеризується параметрами: V_{RAM_j} – об'єм оперативної пам'яті (ГБ); $V_{storage_j}$ – об'єм сховища даних (ГБ); C_j – продуктивність *j*-го вузла (од. навантаження / с); k_{cores_j} – кількість обчислювальних ядер вузла (шт). Кожен з обчислювальних вузлів системи може знаходитись в одному з таких станів: «активний стан» – вузол обробляє певне навантаження та/або здатний прийняти на обробку нові обчислювальні роботи; «стан простою» – вузол є увімкненим, проте не обробляє жодної обчис-

лювальної роботи в даний момент часу; «вимкнений» – вузол виведений з системи шляхом вимикання.

Вхідне навантаження являє собою потік дискретних обчислювальних робіт, що надходять до системи у випадкові моменти часу. Одиницею вхідного навантаження є обчислювальна робота. Компонент роботи називається задачею. Фізично одна робота представляється у вигляді єдиного обчислювального процесу комп'ютера. В процесі розподілу навантаження кожна обчислювальна робота характеризується параметрами: $\Delta t_{\max_i} = const$ (с) – максимальний час виконання роботи (якщо робота не була успішно оброблена по проходженню часу $\Delta t_{\max_i}$, вона виводиться із системи); мінімально необхідний обсяг ресурсів для виконання роботи: мінімально необхідний об'єм оперативної пам'яті $V_{RAM_{\min_i}}$ (ГБ); мінімально необхідна кількість ядер процесора $k_{cores_{\min_i}}$ (шт); мінімально необхідний об'єм постійного сховища даних $V_{storage_{\min_i}}$ (ГБ). Вхідне навантаження характеризується своєю інтенсивністю – обсягом навантаження, що надходить до системи в одиницю часу. Заданою є добова статистика інтенсивності вхідного навантаження на систему.

Моделювання обслуговуючих пристроїв. На відміну від пристрою СМО, обчислювальний вузол (сервер) здатний обробляти одночасно декілька обчислювальних робіт за рахунок розміщення їх на різних ядрах процесора (при чому кількість ядер процесора може бути різною для різних серверів). Крім того, обчислювальна робота (код програми) може бути представлена таким чином, що потребуватиме більше одного обчислювального ядра (за рахунок використання механізмів паралельних обчислень). Ці факти не дозволяють розглядати сервер як обслуговуючий пристрій. Натомість обчислювальне ядро процесора краще підходить для моделювання каналу обслуговування. Можливо вважати, що як і канал СМО, обчислювальне ядро у кожен момент часу знаходиться або у стані "зайняте" або у стані "вільне". Таким чином, йдеться про багатоканальну СМО з різними каналами, адже інтенсивність обробки запитів μ_j обчислювальними ядрами на різних серверах може відрізнятись.

Моделювання вхідного потоку запитів. Розглянемо характеристику вхідного потоку у досліджуваній системі. Характер надходження обчислювальних робіт у загальному випадку залежить від типу роботи, що розглядається. Однак загальна кількість можливих обчислювальних робіт, що потребують обробки у межах розглянутих типів навантаження (SDN, NFV, Edge Computing, Network Slicing і bDDN), є дуже великою. Тому в процесі математичного моделювання пропонується зупинитись на розгляді найбільш загального випадку вхідного потоку зі змінною інтенсивністю та випадковими моментами надходження взаємно незалежних запитів до системи, обслуговування якого становить найбільшу складність через випадковий характер параметрів потоку. Аналіз характеристик і математичне моделювання вхідного потоку для інших типів навантаження пропонується віднести до питань подальших досліджень. За рахунок використання механізмів паралелізації, обчислювальна робота може потребувати для

свого виконання більше ніж один обчислювальний пристрій (ядро процесора). Цей факт не дозволяє напряму ставити у відповідність запиту СМО обчислювальну роботу. Натомість можливим є вважати обчислювальну роботу групою запитів, що надають досліджуваному вхідному потоку запитів властивість неординарності [15]. Водночас неординарним є як процес надходження запитів, так і процес їх обробки, адже обчислювальна робота, що є групою запитів СМО, надходить до системи, направляється на обробку та виводиться із системи лише як єдине ціле.

Таким чином, розглядається вхідний потік із груп запитів, що надходять до системи у випадкові моменти часу зі змінною інтенсивністю надходження. Кількість груп запитів (обчислювальних робіт) є випадковою величиною із дискретним простором значень та неперервним часом. Оцінимо вхідний потік з точки зору характеристик: відсутності післядії, стаціонарності, ординарності.

Оскільки обчислювальні роботи, що розглядаються, не залежать одна від одної, вважатимемо вхідний потік таким, що має властивість відсутності післядії.

Інтенсивність надходження обчислювальних робіт змінюється протягом доби та днів. Так, наприклад, кількість запитів вночі набагато менша, ніж у години найбільшого навантаження, що зазвичай приходиться на вечірні години або середину дня [16]. Тому вхідний потік обчислювальних робіт не має властивості стаціонарності.

Як зазначалось вище, обчислювальна робота являє собою зв'язану групу запитів, що одночасно надходять до системи. Тобто сам потік обчислювальних робіт є ординарним, оскільки імовірність надходження двох і більше обчислювальних робіт від двох користувачів одночасно дуже низька, отже, її можна вважати рівною нулю. Однак вхідний потік запитів СМО не має властивості ординарності. Запити надходять до СМО групами і так само групами обробляються та виводяться із системи.

Таким чином, надалі ми розглядатимемо вхідний потік запитів до досліджуваної системи як *нестаціонарний неординарний (груповий) пуассонівський потік.*

Характеристикою нестаціонарного потоку є його миттєва густина $\lambda(t)$. Миттєвою густиною потоку називається границя відношення середньої кількості подій, що припадає на елементарний відрізок часу $(t, t+\Delta t)$ до довжини цього відрізка, коли остання прямує до нуля [17]:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{m(t+\Delta t) - m(t)}{\Delta t},$$

де $m(t)$ – математичне очікування кількості подій на відрізку часу $(0; t)$.

Для нестаціонарного пуассонівського потоку (тобто ординарного потоку без післядії, але без властивості стаціонарності) кількість подій на інтервалі часу Δt, що розпочався в момент часу t_0, описується законом Пуассона [17]:

$$p_m(\Delta t, t_0) = \frac{a^m \cdot e^{-a}}{m!} (m = 0, 1, 2, ...),$$

де a – математичне очікування кількості подій на інтервалі часу від t_0 до $(t_0+\Delta t)$, що визначається рівністю:

$$a = \int_{t_0}^{t_0+\Delta t} \lambda(t)dt\,.$$

Аналіз цього виразу показує, що a залежить не лише від довжини інтервалу Δt, але й від його положення на осі часу t_0. При моделюванні нестаціонарних потоків необхідно знати закон зміни інтенсивності надходження запитів в часі для періоду, що моделюється. У досліджуваній системі цей закон може бути представлено у вигляді статистичної кривої навантаження на систему протягом доби (рис. 1).

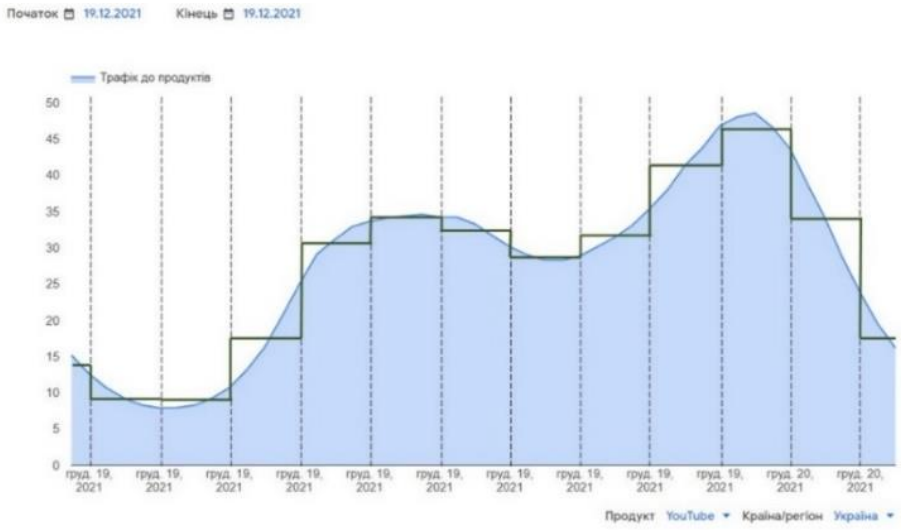


Рис. 1. Приклад представлення $\lambda(t)$ (синя крива) дискретною функцією $\lambda^D(t)$ (зелена ламана) (на прикладі статистики навантаження на відео-сервіс) [16]

Для неординарного вхідного потоку окрім моменту надходження запиту необхідно моделювати також кількість запитів, що можуть одночасно надійти до системи. Кількість одночасних запитів є дискретною випадковою величиною (оскільки обчислювальна робота може виконуватись на $x \in \mathbb{Z}$ ядрах, при чому значення x є обмеженим зверху максимально можливою кількістю ядер процесора одного сервера $X_{\max}$). Для простоти подальших математичних перетворень введемо припущення, що розподіл кількості ядер, на яких повинна виконуватись обчислювальна робота (що одночасно є розподілом кількості запитів неординарного вхідного потоку), у досліджуваній системі є рівномірним. Однак, у загальному випадку, аналогічні розрахунки можуть бути проведені для будь-якого іншого типу розподілу та відповідного йому математичного очікування кількості запитів неординарного вхідного потоку.

Характеристики процесу обслуговування. Обчислювальні роботи спершу надходять до вузла-брокера, який розподіляє роботи за певними правилами між обчис-

лювальними вузлами системи (у найпростішому випадку це правило FIFO – first in first out). Таким чином, на вузлі-брокері формується черга з обчислювальних робіт, що очікують обробки. В загальному випадку ця черга обмежена здатністю апаратного та програмного забезпечення щодо розподілу навантаження. Позначимо цю довжину черги Q. При цьому Q обчислювальних робіт, що містяться у черзі, є групами запитами, кожна з яких складаються з x запитів СМО (у даному випадку, згідно з припущенням, x – це рівномірно розподілена випадкова величина). При переході від обчислювальних робіт до запитів довжина черги стає рівною $Q \cdot M(x)$, де $M(x)$ – математичне очікування випадкової величини x. Таким чином, йдеться про СМО змішаного типу зі скінченною чергою та втратами. Обробка запитів також є груповою, тобто запити надходять в обробку групами по x запитів і так само групами залишають систему. Вважатимемо закон розподілу часу обробки запитів експоненціальним [10]. Інтенсивність обробки кожного каналу обслуговування (обчислювального ядра процесора) в загальному випадку різна та рівна μ_j.

Таким чином, математичною моделлю процесу розподіленого обслуговування навантаження у досліджуваній системі можна вважати багатоканальну СМО зі скінченною чергою довжини $Q \cdot M(x)$ та з втратами. Вхідний потік запитів є нестаціонарним неординарним потоком зі змінною інтенсивністю $\lambda(t)$ та випадковою величиною кількості запитів x, що одночасно надходять до системи. Обслуговуючі пристрої представлені ядрами процесорів у системі та мають різну інтенсивність обробки μ_j. Обчислювальна робота формує групу запитів. Кожен запит займає один обслуговуючий пристрій. При цьому потік обчислювальних робіт є ординарним (на відміну від потоку запитів).

Через наявність властивостей нестаціонарності та неординарності визначити показники ефективності такої системи (середній час в обробці, ймовірність втрати запиту тощо) за стандартними формулами для найпростішого вхідного потоку не вдається. Тому з метою переходу до спрощених розрахунків пропонується така методика перетворення математичної моделі:

- перехід до стаціонарного неординарного потоку на підставі представлення кривої інтенсивності надходження обчислювальних робіт до системи у вигляді дискретної функції;

- застосування переходу до комплектів серверів для визначення втрат при неординарному потоці заявок.

У процесі спрощення моделі особливу увагу приділено збереженню її точності та адекватності реальному об'єкту дослідження. Водночас свідомо допускаються певні втрати точності, але зазначаються параметри трансформованої моделі, що впливають на її точність.

Інтенсивність вхідного потоку запитів є функцією часу $\lambda(t)$ і в загальному випадку ця функція є неперервною. Однак із певним наближенням можна представити цю функцію у вигляді дискретної функції $\lambda^D(t)$, при чому величина кроку дискрети-

зації впливатиме на точність наближених розрахунків та адекватність моделі реальному об'єктові відповідно (рис. 1). Для визначення оптимальної величини кроку дискретизації пропонується використання методу квантування за рівнями [18]. Це дозволяє узгодити величину кроку дискретизації функції зі швидкістю зміни інтенсивності вхідного навантаження. Для визначення кроку квантування пропонується метод розрахунку порогових величин інтенсивностей вхідного навантаження $\lambda_n(t)$ з урахуванням кількості обчислювальних вузлів у системі за формулою:

$$\lambda_n(t) = \mu' \cdot k' + Q \sqrt{\frac{p_{loss} \cdot k'^{Q} \cdot k'!}{p_0}}, k' \in \mathbb{N}^*, \quad (1)$$

де μ' – середня інтенсивність обробки запитів; k' – кількість обслуговуючих каналів; p_{loss} – задана імовірність втрати запиту; p_0 – імовірність знаходження системи у нульовому стані.

Таким чином, надалі ми можемо розглядати стаціонарний неординарний вхідний потік на кожному з інтервалів Δt, інтенсивність якого дорівнює середньому значенню функції $\lambda(t)$ на відповідному відрізку, а модель є значно простішою. При чому розмір інтервалів дискретизації визначається моментами переходу між сусідніми $\lambda_n(t)$. Особливості роботи зі стаціонарним неординарним потоком було розглянуто зокрема у роботі [19], де було запропоновано перейти від неординарного потоку до ординарного шляхом переходу до розгляду комплектів каналів (пристроїв обслуговування), кожен з яких обробляє один груповий запит. При сталій кількості запитів у групі такий перехід здійснюється шляхом ділення кількості серверів на кількість запитів у групі. Оскільки у досліджуваній системі кількість запитів у груповому потоці не є сталою, при переході до комплектів серверів доцільно розглядати математичне очікування випадкової величини як дільник для перетвореної кількості серверів. Таким чином, якщо у досліджуваній системі S кількість каналів (ядер процесора) була рівною k, то у перетвореній системі S' кількість каналів буде рівною $k' = k / M(x)$, де $M(x)$ – математичне очікування кількості одночасних запитів x у групі. Отже, отримуємо перетворену СМО, в якій вхідний потік запитів є ординарним та стаціонарним на кожному із проміжків часу Δt. По суті, перетворена система S' являє собою систему із запитами, що відповідають нібито "однаковим обчислювальним роботам", які розпаралелені для виконання на $M(x)$ ядрах. Обчислювальними каналами в такій перетвореній системі є уявні сервери з однаковою кількістю ядер $M(x)$. Оскільки визначено інтенсивність надходження запитів на інтервалі Δt як $\lambda^t = \lambda^D(t \mid t \in (t_0, t_0 + \Delta t])$, то при переході до перетвореної системи S' отримуємо інтенсивність надходження відповідних комплектів запитів рівну $\lambda' = \frac{\lambda^t}{M(x)} = \frac{\int_{t_0}^{t_0+\Delta t} \lambda(t)dt}{\Delta t \cdot M(x)}$.

Довжина черги у перетвореній системі є рівною $Q' = \frac{Q \cdot M(x)}{M(x)} = Q$, що статистично співпадає із довжиною черги обчислювальних робіт. Потік вхідних запитів у такій системі є пуассонівським, а розподіл часу обробки запитів – експоненціальним.

Каналами обслуговування виступають k' груп з $M(x) = \frac{X_{\max} + X_{\min}}{2}$ ядер процесора (у випадку рівномірного розподілу x), де $X_{\max}$ – максимальна кількість ядер процесора одного сервера у досліджуваній системі (що відповідає максимально можливій кількості ядер, яку може вимагати обчислювальна робота для обробки); $X_{\min} = 1$ – мінімально можлива кількість ядер, яку може вимагати обчислювальна робота для обробки. При переході до системи S', сумарна інтенсивність i-го комплекту каналів відповідно буде дорівнювати $\mu_i' = \sum_{j=1}^{M(x)} \mu_j$ запитів в одиницю часу. В загальному випадку інтенсивність обробки μ_j для кожного обчислювального ядра є різною та визначається фізичними параметрами ядра. Однак під час розгляду комплектів ядер сумарну інтенсивність кожного комплекту із деякою похибкою можна вважати сталою та такою, що дорівнює середньому значенню суми інтенсивностей у межах кожного комплекту:

$$\mu' = \frac{\sum_{i=1}^{k'} \mu_i'}{k'} = \frac{\sum_{i=1}^{k'} \sum_{j=1}^{M(x)} \mu_{ij}}{k'},$$

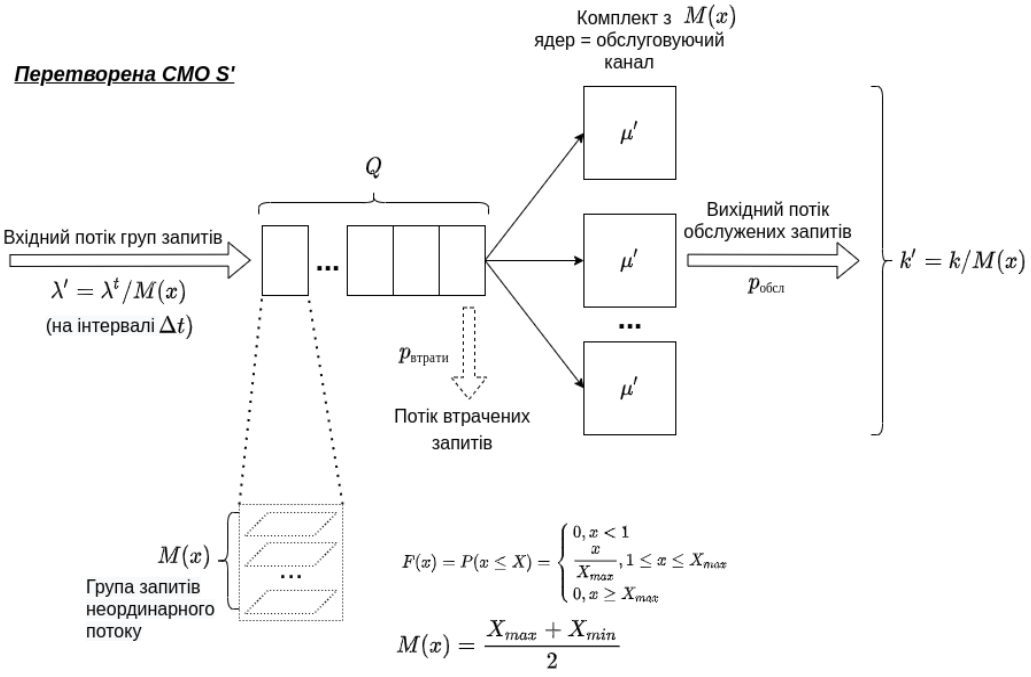
де μ_{ij} – середня інтенсивність обробки для j-го ядра i-го комплекту каналів. Графічне представлення перетвореної СМО S' наведено на рисунку (рис. 2, а).

Аналізуючи отриману систему, можливо розглядати деяку скінченну кількість станів СМО S', при чому перехід з одного стану в інший здійснюється під дією подій надходження та обробки запитів. Потоки надходження та обробки запитів у перетвореній системі S' на інтервалі часу Δt можна вважати найпростішими, що дає змогу представити цю систему у вигляді марківського процесу. Граф станів такого процесу зображено на рис. 2, б, де використовується така множина станів: S_0 – в системі немає жодного запиту, всі канали вільні, черги немає; S_1 – в системі 1 запит, 1 канал зайнятий відповідно, черги немає;…; $S_{k'}$ – в системі k' запитів, всі канали зайняті, черги немає; $S_{k'+1}$ – в системі $k'+1$ запит, всі канали зайняті, один запит очікує в черзі;...; $S_{k'+Q}$ – в системі $k'+Q$ запитів, всі канали та місця в черзі зайняті.

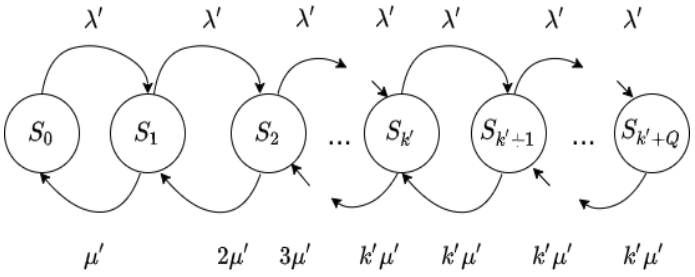
Склавши та розв'язавши систему диференційних рівнянь Колмогорова для ймовірностей станів такої системи, позначивши для зручності $\rho = \lambda'/\mu'$, отримуємо

значення для граничних імовірностей: $p_0 = [1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + ... + \frac{\rho^{k'}}{k'!} + ... + \frac{\rho^{k'+1} \cdot (1 - (\frac{\rho}{k'})^Q)}{k' \cdot k'! \cdot (1 - \frac{\rho}{k'})}]^{-1}$;

$p_j = \frac{\rho^j}{j!} \cdot p_0, (j = 1, ..., k')$; $p_{k'+r} = \frac{\rho^{k'+r}}{k'^r \cdot k'!} \cdot p_0, (r = 1, ..., Q)$.



а)



б)

Рис. 2. Графічне представлення (а) та граф станів (б) для перетвореної СМО S'

Таким чином, отримано математичну модель системи розподіленого обслуговування навантаження ІКМ у вигляді СМО зі скінченною чергою та з втратами, що описується такими параметрами:

$k' = k / M(x)$ – кількість обслуговуючих каналів, де $k = k_{cores}$ – задана за умовою кількість обчислювальних ядер у системі;

$M(x)$ – математичне очікування кількості одночасних запитів x у групі;

$\lambda' = \frac{\int_{t_0}^{t_0+\Delta t} \lambda(t)dt}{\Delta t \cdot M(x)}$ – інтенсивність надходження запитів на відрізку часу $[t_0; t_0 + \Delta t]$, де $\lambda(t)$ – задана за умовою крива вхідного навантаження;

$\mu' = \frac{\sum_{i=1}^{k'} \sum_{j=1}^{M(x)} \mu_{ij}}{k'}$ – середня інтенсивність обробки заявок, де $\mu_{ij} = \overline{\mu_j}$ – середня інтенсивність обробки *j*-го обчислювального ядра.

$Q' = \frac{Q \cdot M(x)}{M(x)} = Q$ – довжина черги (співпадає із заданою довжиною черги).

Основними показниками ефективності СМО зі скінченною чергою та з втратами, формальні вирази для яких можна отримати із формул Літтла [20], є:

- ймовірність втрати запиту: $p_{loss} = p_{k'+Q} = \frac{\rho^{k'+Q}}{k'^{Q} \cdot k'!} p_0$;

- абсолютна пропускна здатність: $A = \lambda'(1 - \frac{\rho^{k'+Q}}{k'^{Q} \cdot k'!} p_0)$;

- середня кількість заявок у черзі: $L_Q = \frac{\rho^{k'+1} \cdot p_0 (1 - (Q + 1 - Q\frac{\rho}{k'}) \cdot (\frac{\rho}{k'})^Q}{k' \cdot k'! (1 - \frac{\rho}{k'})^2}$;

- середня кількість заявок в обробці, що відповідає середній кількості зайнятих каналів обслуговування: $\bar{n} = \rho \cdot (1 - \frac{\rho^{k'+Q}}{k'^{Q} \cdot k'!} p_0)$, де $\rho = \lambda' / \mu'$.

III. Оцінка адекватності запропонованої моделі

На основі побудованої математичної моделі було запропоновано комплексний метод енергоефективного обслуговування навантаження ІКМ, докладно описаний у роботі [21]. Побудована математична модель у вигляді СМО лежить в основі кроку 2 «Визначення шаблонів горизонтального масштабування» та дозволяє визначати оптимальну кількість активних обчислювальних вузлів у системі на кожному інтервалі часу, що визначається швидкістю зміни інтенсивності вхідного навантаження, за умови гарантування виконання вимог щодо обслуговування навантаження ІКМ, зокрема вимог щодо доступності системи та продуктивності обслуговування, що безпосередньо впливає на час відповіді на запит. Для визначення кількості необхідних активних обчислювальних вузлів застосовується формула (1). Перевірку ефективності комплексного методу енергоефективного обслуговування навантаження ІКМ на основі побудованої СМО було здійснено методом лабораторного експерименту, що докладно описано у роботі [22]. За результатами експериментальної перевірки запропонований комплексний метод забезпечує виконання вимог щодо доступності системи обслуговування та дає виграш за запропонованим критерієм ефективності $K_{opt} = E_\Sigma \cdot C_\Sigma \to \max$, при

$p_{avail} \geq p_{avail_{SLA}}$, де $E_\Sigma = \dfrac{\omega}{\sum_{j=1}^{N} W_j}$ – показник енергоефективності; $C_\Sigma = \dfrac{\omega}{T}$ – показник продуктивності обслуговування навантаження; ω – обсяг корисного навантаження, що обробляється системою; T – час обробки; $W = \sum_{j=1}^{N} W_j$ – сумарний обсяг спожитої енергії при обробці навантаження на N вузлах. Виграш за запропонованим критерієм складає 15,722% у порівнянні із відомим енергоефективним підходом Backfill [23] та 88,887% у порівнянні з широко використовуваним підходом Round Robin [24].

Висновки

1. В результаті проведених досліджень запропоновано математичну модель системи розподіленого обслуговування навантаження ІКМ у вигляді СМО, яка відрізняється від відомих тим, що враховує змінний характер інтенсивності вхідного навантаження та можливість використання технік паралелізації при розробці ПЗ сучасних ІКМ. Це дозволило отримати кількісний опис взаємозв'язків між параметрами системи обслуговування та показниками якості обслуговування навантаження ІКМ.
2. В процесі математичного моделювання системи розподіленого обслуговування навантаження запропоновано метод переходу від нестаціонарного неординарного вхідного потоку заявок до стаціонарного ординарного потоку шляхом дискретизації кривої інтенсивності вхідного навантаження та за допомогою переходу до комплектів серверів. Для дискретизації кривої інтенсивності вхідного навантаження запропоновано використання методу квантування за рівнями, що дозволило узгодити величину кроку дискретизації функції зі швидкістю зміни інтенсивності вхідного навантаження. Можливі втрати точності моделі в процесі спрощення регулюються параметрами трансформованої моделі, що дозволяє зберегти адекватність моделі об'єкта дослідження.
3. У результаті математичного моделювання отримано математичні вирази для показників ефективності СМО: імовірності втрати запиту у змодельованій системі, її абсолютної пропускної здатності, середньої кількості заявок у черзі та обробці.
4. На основі побудованої СМО розроблено комплексний метод енергоефективного обслуговування навантаження [21]. Адекватність моделі та ефективність методу перевірено за допомогою лабораторного експерименту. Виграш за запропонованим критерієм ефективності склав 15,722% та 88,887% у порівнянні з широко використовуваними підходами Backfill і Round Robin відповідно.
5. Перспектива подальших досліджень у цій області полягає в більш докладному аналізі окремих типів навантаження ІКМ і відповідних параметрів вхідного потоку СМО, сформованого цими типами навантаження. Крім того, заслуговує подальшого дослідження тема змішаних потоків навантаження, пріоритезації запитів та їх розподілу між обчислювальними вузлами на основі типізації обчислювального навантаження.

Список літератури

1. ITU-T Recommendation Y.3300, (2014), "Framework of software-defined networking", available at: https://www.itu.int/rec/T-REC-Y.3300-201406-I/en (last accessed 16.09.2022)

2. ETSI Standard GR NFV 003, (2020), "Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV", available at: https://www.etsi.org/deliver/etsi_gr/NFV/001_099/003/01.05.01_60/gr_NFV003v010501p.pdf (last accessed 16.09.2022)

3. *Khan, L.U., Yaqoob, I., Tran, N.H., Han, Z. and Hong, C.S.* (2020), "Network slicing: Recent advances, taxonomy, requirements, and open research challenges", IEEE Access, No. 8, P. 36009-36028. **DOI**: https://doi.org/10.1109/ACCESS.2020.2975072

4. *Bravo, C., Bäckström, H.* (2020), "Edge computing and deployment strategies for communication service providers", White Paper, Ericsson, available at: https://www.ericsson.com/en/reports-and-papers/white-papers/edge-computing-and-deployment-strategies-forcommunication-service-providers (last accessed 10.04.2022)

5. ITU-T Recommendation Y.3652, (2020), "Big data driven networking - requirements", available at: https://www.itu.int/rec/T-REC-Y.3652 (last accessed 16.09.2022)

6. ITU-T Recommendation Y.3156, (2020), "Framework of network slicing with AI-assisted analysis in IMT-2020 networks", available at: https://www.itu.int/rec/T-REC-Y.3156/en (last accessed 16.09.2022)

7. ITU-T Recommendation F.743.12, (2021), "Requirements for edge computing in video surveillance", available at: https://www.itu.int/rec/T-REC-F.743.12/en (last accessed 16.09.2022)

8. *Redana, S., Bulakci, Ö., Zafeiropoulos, A., Gavras, A., Tzanakaki, A., Albanese, A., Kousaridas, A., Weit, A., Sayadi, B., Jou, B.T., Bernardos, C.J.* (2019), "5G PPP architecture working group: View on 5G architecture", Brussels, Belgium: European Commission, 2019, 182 p., **DOI**: http://doi.org/10.5281/zenodo.3265031

9. *Carvalho, G.H., Woungang, I., Anpalagan, A. and Jaseemuddin, M.* (2018), "Analysis of joint parallelism in wireless and cloud domains on mobile edge computing over 5G systems", Journal of Communications and Networks, No. 20(6), P. 565-577. **DOI**: https://doi.org/10.1109/JCN.2018.000089

10. *Bai, W.H., Xi, J.Q., Zhu, J.X., Huang, S.W.* (2015), "Performance analysis of heterogeneous data centers in cloud computing using a complex queuing model", Mathematical Problems in Engineering, No. 2015, P. 1-15. **DOI**: https://doi.org/10.1155/2015/980945

11. *Van der Boor, M., Comte, C.* (2021), "Load balancing in heterogeneous server clusters: Insights from a product-form queueing model", Proceedings of the 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), P. 1-10. **DOI**: https://doi.org/10.1109/IWQOS52092.2021.9521355

12. *Hu, C., Deng, Y., Yang, L.T., Zhao, Y.* (2019), "Estimating the resource demand in power-aware clusters by regressing a linearly dependent relation", IEEE Transactions on Sustainable Computing, No. 6(3), P. 385-397. **DOI**: https://doi.org/10.1109/TSUSC.2019.2894708

13. *Meisner, D., Wenisch, T.F.* (2010), "Stochastic queuing simulation for data center workloads", Proceedings of the Exascale Evaluation and Research Techniques Workshop, No. 16, p. 1-9.

14. *Lee, J., Nam, B.G., Yoo, H.J.* (2007), "Dynamic voltage and frequency scaling (DVFS) scheme for multi-domains power management", Proceedings of the 2007 IEEE Asian Solid-State Circuits Conference, P. 360-363. **DOI**: https://doi.org/10.1109/ASSCC.2007.4425705

15. *Litvinov, A.L.* (2018), Theory of queuing systems: training manual, O.M. Beketov NUUEKh, Kharkiv, 141 p. [*Литвинов, А.Л.* (2018), Теорія систем масового обслуговування: навч. посібник, ХНУМГ ім. О. М. Бекетова, Харків, 141 с.]

16. Google Transparency Report, (2021), available at: https://transparencyreport.google.com/traffic/overview (last accessed 23.12.2021)

17. *Wentzel, E.S.* Probability Theory (4th ed.), M.: Nauka, 1969, 576 p. [Вентцель, Е.С. *Теория вероятностей (4-е изд.)*, М.: Наука, 1969, 576 с.]

18. *Naumov, E.V., Berezin, V.B., Berezin, V.V., Gataulin, V.M., Monchak, A.M.* (2004), "Implementation of analog-to-digital conversion of optical signals with a variable step", Izvestia of higher educational institutions of Russia, Radioelectronics, No. 3, P. 57-65. [Наумов, Е.В., Березин, В.Б., Березин, В.В., Гатаулин, В.М., Мончак, А.М. (2004), "Реализация аналого-цифрового преобразования оптических сигналов с переменным шагом", Известия высших учебных заведений России, Радиоэлектроника, No. 3, С. 57-65.]

19. *Ложковский, А.Г.* (2012), "Теория массового обслуживания в телекоммуникациях: підручник", Одесса: ОНАЗ ім. ОС Попова, 112 с.

20. *Kremer, N.S., Putko, B.A., Trishyn, I.M., Fridman, M.N.* (2017), "Research of operations in economics: a textbook for universities", Ed. Yurayt, Moscow, 438 p. [Кремер, Н.Ш., Путко, Б.А., Тришин, И.М., Фридман, М.Н. (2017), "Исследование операций в экономике: учебник для вузов", Издательство Юрайт, Москва, 438 с.]

21. *Globa, L., Gvozdetska, N., Prokopets, V.* (2021), "Providing Energy-efficient and High-performance Infrastructure for Smart Network", Proceedings of the 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), P. 133-136, **DOI**: https://doi.org/10.1109/UkrMiCo52950.2021.9716620

22. *Globa, L., Gvozdetska, N.* (2021), "Experimental analysis of PCPB-2: Comprehensive energy-efficient approach to distributed workload processing in communication networks", Proceedings of the 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), P. 1-3. **DOI**: https://doi.org/10.1109/BlackSeaCom52164.2021.9527759

23. IBM Spectrum LSF 10.1.0 Document, (2022), available at: https://www.ibm.com/docs/en/spectrum-lsf/10.1.0?topic=jobs-backfill-scheduling (last accessed 04.07.2022)

24. *Vashistha, J., Jayswal, A.K.* (2013), "Comparative study of load balancing algorithms", IOSR Journal of Engineering, No. 3(3), P. 45-50.